

# Global Relation Embedding for Relation Extraction

Yu Su\*, Honglei Liu\*, Semih Yavuz, Izzeddin Gür

University of California, Santa Barbara

{ysu, honglei, syavuz, izzeddinur}@cs.ucsb.edu

Huan Sun

The Ohio State University

sun.397@osu.edu

Xifeng Yan

University of California, Santa Barbara

xyan@cs.ucsb.edu

## Abstract

Recent studies have shown that embedding textual relations using deep neural networks greatly helps relation extraction. However, many existing studies rely on supervised learning; their performance is dramatically limited by the availability of training data. In this work, we generalize textual relation embedding to the distant supervision setting, where much larger-scale but noisy training data is available. We propose leveraging *global statistics* of relations, i.e., the co-occurrence statistics of textual and knowledge base relations collected from the entire corpus, to embed textual relations. This approach turns out to be more robust to the training noise introduced by distant supervision. On a popular relation extraction dataset, we show that the learned textual relation embeddings can be used to augment existing relation extraction models and significantly improve their performance. Most remarkably, for the top 1,000 relational facts discovered by the best existing model, the precision can be improved from 83.9% to 89.3%.

## 1 Introduction

Relation extraction requires deep understanding of the relation between entities. Early studies mainly use hand-crafted features (Kambhatla, 2004; GuoDong et al., 2005), and later kernel methods are introduced to automatically generate features (Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Zhang et al., 2006). Recently neural network models have been introduced to embed words, relations, and

sentences in continuous feature space, and have shown a remarkable success in relation extraction (Socher et al., 2012; Zeng et al., 2014; Xu et al., 2015b; Zeng et al., 2015; Lin et al., 2016). In this work, we study the problem of embedding *textual relations*, defined as the shortest dependency path<sup>1</sup> between two entities in the dependency graph of a sentence, to improve relation extraction.

Textual relations are one of the most discriminative textual signals that lay the foundation of many relation extraction models. Because of their exact feature matching, early kernel based models (Bunescu and Mooney, 2005) can hardly exploit fine-grained word similarities. More recent studies (Xu et al., 2015a,b, 2016; Liu et al., 2016) have explored embedding textual relations via neural networks. However, they have all focused on the *supervised* setting, where the embedding model is trained on a set of sentences with manually annotated target relation. The high cost of manual annotation limits the scale of their setting: The training data typically consists of several thousands of annotated sentences and around 10 target relations (Liu et al., 2016).

In contrast, we embed textual relations with *distant supervision* (Mintz et al., 2009), which provides much larger-scale training data without the need of manual annotation. However, the assertion of distant supervision, “any sentence containing a pair of entities that participate in a knowledge base (KB) relation is likely to express the relation,” can be violated more often than not, resulting in many wrongly labeled training examples. A representative example is shown in Figure 1. Embedding quality is thus compromised by the noise in training data.

Instead of using *local statistics*, i.e., individual

<sup>1</sup>We use fully lexicalized shortest dependency path with directional and typed dependency relations.

\* Equally contributed.

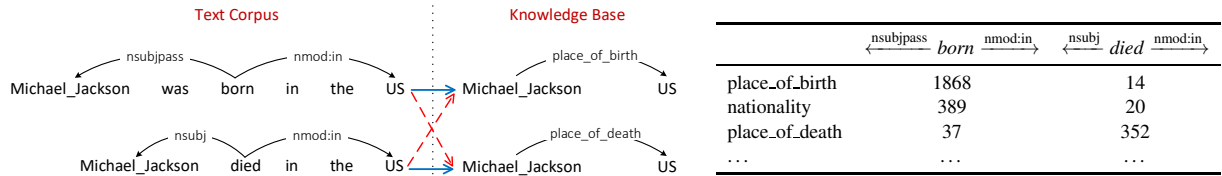


Figure 1: The wrong labeling problem of distant supervision, and how to combat it with global statistics. *Left*: conventional distant supervision. Each of the textual relations will be labeled with both KB relations, while only one is correct (blue and solid), and the other is wrong (red and dashed). *Right*: distant supervision with global statistics. The two textual relations can be clearly distinguished by their co-occurrence distribution of KB relations. Statistics are based on the annotated ClueWeb data released in (Toutanova et al., 2015).

textual-KB relation pairs, like in previous work, we propose to embed textual relations using *global statistics*, which provide a natural solution to the wrong labeling problem. More specifically, we collect *co-occurrence* statistics of textual and KB relations from the entire corpus, where the co-occurring relationship is established via distant supervision. The semantics of a textual relation can then be represented by its co-occurrence distribution of KB relations. For example, the distribution in Figure 1 indicates that the textual relation  $\text{SUBJECT} \xleftarrow{\text{nsubjpass}} \text{born} \xrightarrow{\text{nmod:in}} \text{OBJECT}$  mostly means *place\_of\_birth*, and is also a good indicator of *nationality*, but not *place\_of\_death*. Textual relation embeddings learned on such global statistics are thus more robust to the noise introduced by the wrong labeling problem.

We augment existing relation extractions using the learned textual relation embeddings. On a popular dataset introduced by Riedel et al. (2010), we show that a number of recent relation extraction models, which are based on local statistics, can be significantly improved using our textual relation embeddings. Most remarkably, a new best performance is achieved when augmenting the best existing model with our relation embeddings: The precision of the top 1,000 relational facts discovered by the model is improved from 83.9% to 89.3%, a 33.5% decrease in error rate. The results suggest that relation embedding with global statistics can capture complementary information to existing local statistics based models.

## 2 Related Work

Relation extraction is an important task in information extraction, and has attracted substantial attention. Early relation extraction methods are mainly feature-based (Kambhatla, 2004; GuoDong et al., 2005), where features at various levels, including POS tags, constituency and dependency parses, are integrated in a max entropy

model. With the popularity of kernel methods, a large number of kernel-based relation extraction models have been proposed (Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Zhang et al., 2006). The most related work to ours is by Bunescu and Mooney (2005), where the authors point out and demonstrate the importance of shortest dependency paths for relation extraction.

More recently, the focus of relation extraction research has been revolving around neural network models, which can alleviate the problem of exact feature matching of previous methods and have shown a remarkable success (e.g., (Socher et al., 2012; Zeng et al., 2014)). Among those, the most related are the ones embedding shortest dependency paths with neural networks (Xu et al., 2015a,b, 2016; Liu et al., 2016). For example, Xu et al. (2015b) use a recurrent neural network (RNN) with LSTM units to embed shortest dependency paths without typed dependency relations, while a convolutional neural network is used in (Xu et al., 2015a). However, they are all based on the supervised setting with a limited scale.

Distant supervision (Mintz et al., 2009) has emerged as an appealing way to solicit large-scale training data for relation extraction. Various efforts have been put to combat the long-criticized wrong labeling problem. Riedel et al. (2010), Hoffmann et al. (2011), and Surdeanu et al. (2012) have attempted a multi-instance learning (Dietterich et al., 1997) framework to soften the assumption of distant supervision, but their models are still feature-based. Zeng et al. (2015) combine multi-instance learning with neural networks, with the assumption that at least one of the contextual sentences of an entity pair is expressing the target relation, but this will lose useful information in the neglected sentences. Instead, Lin et al. (2016) use all the contextual sentences, and

introduce an attentive neural network to learn to properly weight the contextual sentences. However, no prior study has exploited global statistics to combat the wrong labeling problem of distant supervision.

In universal schema (Riedel et al., 2013) for KB completion and relation extraction as well as its extensions (Toutanova et al., 2015; Verga et al., 2016), a binary matrix is constructed from the entire corpus, with entity pairs as rows and textual/KB relations as columns. A matrix entry is 1 if the relational fact is observed in training, and 0 otherwise. Entity pair and relation embeddings, either directly or via neural networks, are then learned on the matrix entries, which are still individual relational facts, and the wrong labeling problem remains. Global co-occurrence frequencies are not taken into account, which is the focus of this study.

### 3 Global Statistics of Relations

When using a corpus to train statistical models, there are two levels of statistics to exploit: *local* and *global*. Take word embedding as an example. The skip-gram model (Mikolov et al., 2013) is based on local statistics: During training, we sweep through the corpus and slightly tune the embedding model based on each local window (e.g., 10 consecutive words). In contrast, in global statistics based methods, exemplified by latent semantic analysis (Deerwester et al., 1990) and GloVe (Pennington et al., 2014), we process the entire corpus to collect global statistics like word-word co-occurrence counts, normalize the raw statistics, and train an embedding model directly on the normalized global statistics.

Most existing studies on relation extraction are based on local statistics of relations, i.e., models are trained on individual relation examples. In this section, we describe how we collect co-occurrence statistics of textual and KB relations, and how to normalize the raw statistics. By the end of this section a bipartite *relation graph* like Figure 2 will be constructed, with one node set being textual relations  $\mathcal{T}$ , and the other being KB relations  $\mathcal{R}$ . The edges are weighted by the normalized co-occurrence statistics of relations.

#### 3.1 Relation Graph Construction

Given a corpus and a KB, we first do entity linking on each sentence, and do dependency pars-

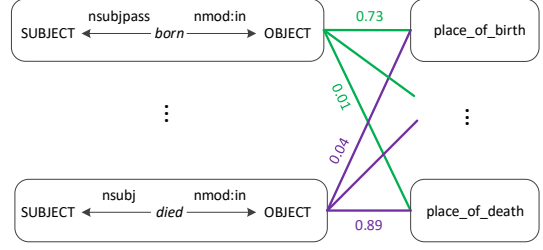


Figure 2: Relation graph. The left node set is textual relations, and the right node set is KB relations. The raw co-occurrence counts are normalized such that the KB relations corresponding to the same textual relation form a valid probability distribution. Edges are colored by textual relation and weighted by normalized co-occurrence statistics.

ing if at least two entities are identified<sup>2</sup>. For each entity pair  $(e, e')$  in the sentence, we extract the fully lexicalized shortest dependency path as a textual relation  $t$ , forming a *relational fact*  $(e, t, e')$ . There are two outcomes from this step: a set of textual relations  $\mathcal{T} = \{t_i\}$ , and the *support*  $S(t_i)$  for each  $t_i$ . The support of a textual relation is a *multiset* containing the entity pairs of the textual relation. The *multiplicity* of an entity pair,  $m_{S(t_i)}(e, e')$ , is the number of occurrences of the corresponding relational fact  $(e, t_i, e')$  in the corpus. For example, if the support of  $t_i$  is  $S(t_i) = \{(e_1, e'_1), (e_1, e'_1), (e_2, e'_2), \dots\}$ , entity pair  $(e_1, e'_1)$  has a multiplicity of 2 because the relational fact  $(e_1, t_i, e'_1)$  occur in two sentences. We also get a set of KB relations  $\mathcal{R} = \{r_j\}$ , and the support  $S(r_j)$  of a KB relation  $r_j$  is the set of entity pairs having this relation in the KB, i.e., there is a relational fact  $(e, r_j, e')$  in the KB. The number of *co-occurrences* of a textual relation  $t_i$  and a KB relation  $r_j$  is

$$n_{ij} = \sum_{(e, e') \in S(r_j)} m_{S(t_i)}(e, e'), \quad (1)$$

i.e., every occurrence of the relational fact  $(e, t_i, e')$  will be counted as a co-occurrence of  $t_i$  and  $r_j$  if  $(e, e') \in S(r_j)$ . A bipartite relation graph can then be constructed, with  $\mathcal{T}$  and  $\mathcal{R}$  as the node sets, and the edge between  $t_i$  and  $r_j$  has weight  $n_{ij}$  (no edge if  $n_{ij} = 0$ ), which will be normalized later.

<sup>2</sup>In the experiments entity linking is assumed given, and dependency parsing is done using Stanford Parser (Chen and Manning, 2014) with universal dependencies.

### 3.2 Normalization

The raw co-occurrence counts have a heavily skewed distribution that spans several orders of magnitude: A small portion of relation pairs co-occur highly frequently, while most relation pairs co-occur only a few times. For example, a textual relation,  $\text{SUBJECT} \xleftarrow{\text{nsubjpass}} \text{born} \xrightarrow{\text{nmod:in}} \text{OBJECT}$ , may co-occur with the KB relation `place_of_birth` thousands of times (e.g., “Michelle Obama was born in Chicago”), while a synonymous but slightly more compositional textual relation,  $\text{SUBJECT} \xleftarrow{\text{nsubjpass}} \text{born} \xrightarrow{\text{nmod:in}} \text{city} \xrightarrow{\text{nmod:of}} \text{OBJECT}$ , may co-occur with the KB relation only several times in the whole corpus (e.g., “Michelle Obama was born in the city of Chicago”). Learning directly on the raw co-occurrence counts, an embedding model may put a disproportionate amount of weight on the most frequent relations, and may not learn well on the majority of rarer relations. Proper normalization is therefore necessary.

A number of normalization strategies have been proposed in the context of word embedding, including correlation- and entropy-based normalization (Rohde et al., 2005), positive pointwise mutual information (Bullinaria and Levy, 2007), and some square root type transformation (Lebret and Collobert, 2014). A shared goal is to reduce the impact of the most frequent words, e.g., “the” and “is,” which tend to be less informative for the purpose of embedding.

We have a similar goal, but from preliminary studies we find that a somewhat more aggressive normalization strategy works better for relations: for each textual relation, we normalize its co-occurrence counts to form a probability distribution over KB relations. The new edge weights of the relation graph thus become  $w_{ij} = \tilde{p}(r_j|t_i) = n_{ij} / \sum_{j'} n_{ij'}$ . Compared with the existing normalization strategies for word embedding, this will give more weights to the rare relations. Every textual relation now is associated with a set of edges whose weights sum to 1. It can be justified by the different distribution of words and relations. Because textual relations are sequences, they are less likely to collide than individual words, and rare relations are the norm, not the exception (a comparative analysis can be found in Appendix A). Therefore, we need to focus more on the rare relations.

## 4 Textual Relation Embedding

Next we discuss how to learn embedding of textual relations based on the constructed relation graph. We call our approach **Global Relation Embedding (GloRE)** in light of global statistics of relations.

### 4.1 Embedding via RNN

Given the relation graph, a straightforward way of relation embedding is matrix factorization, similar to latent semantic analysis (Deerwester et al., 1990) for word embedding. However, textual relations are different from words in that they are sequences composed of words and typed dependency relations. Therefore, we use recurrent neural networks (RNNs) for embedding, which respect the compositionality of textual relations and can learn the shared sub-structures of different textual relations (Toutanova et al., 2015). For the examples in Figure 1, an RNN can learn, from both textual relations, that the shared dependency relation “nmod:in” is indicative of location modifiers.

For a textual relation, we first decompose it into a sequence of tokens  $\{x_1, \dots, x_m\}$ , which includes lexical words and directional dependency relations. For example, the textual relation  $\text{SUBJECT} \xleftarrow{\text{nsubjpass}} \text{born} \xrightarrow{\text{nmod:in}} \text{OBJECT}$  is decomposed into a sequence of three tokens  $\{-\text{nsubjpass}, \text{born}, \text{nmod:in}\}$ , where “-” represents a left arrow. Note that we include directional dependency relations, because both the relation type and the direction are critical in determining the meaning of a textual relation. For example, the dependency relation “nmod:in” often indicates a location modifier and is thus strongly associated with location-related KB relations like `place_of_birth`. The direction also plays an important role. Without knowing the direction of the dependency relations, it is impossible to distinguish `child_of` and `parent_of`.

An RNN with gated recurrent units (GRUs) (Cho et al., 2014) is then applied to consecutively process the sequence as shown in Figure 3. We have also explored more advanced constructs like attention, but the results are similar, so we opt for a vanilla RNN in consideration of model simplicity.

Let  $\phi$  denote the function that maps a token  $x_l$  to a fixed-dimensional vector, the hidden state vectors of the RNN are calculated recursively:

$$h_l = \text{GRU}(\phi(x_l), h_{l-1}). \quad (2)$$

GRU follows the definition in Cho et al. (2014):



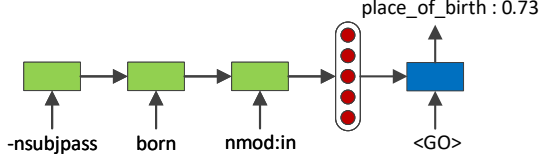


Figure 3: Embedding model. *Left*: A RNN with GRU for embedding. *Middle*: embedding of textual relation. *Right*: a separate GRU cell to map a textual relation embedding to a probability distribution over KB relations.

$$\begin{aligned}
z_l &= \sigma(W_z \phi(x_l) + U_z h_{l-1}) \\
r_l &= \sigma(W_r \phi(x_l) + U_r h_{l-1}) \\
\tilde{h}_l &= \tanh(W_h \phi(x_l) + U_h(r_l \circ h_{l-1})) \\
h_l &= z_l \circ h_{l-1} + (1 - z_l) \circ \tilde{h}_l
\end{aligned}$$

where  $\sigma$  is the sigmoid function, the operator  $\circ$  is element-wise multiplication,  $W_*$  and  $U_*$  are parameter matrices of the model,  $z_l$  is the update gate vector and  $r_l$  is the reset gate vector. We use the final hidden state vector  $h_m$  as the embedding of the textual relation.

## 4.2 Training Objective

We use global statistics in the relation graph to train the embedding model. Specifically, we model the semantics of a textual relation as its co-occurrence distribution of KB relations, and learn textual relation embeddings to reconstruct the corresponding co-occurrence distributions.

We use a separate GRU cell followed by softmax to map a textual relation embedding to a distribution over KB relations (Figure 3); the full model thus resembles the sequence-to-sequence architecture (Sutskever et al., 2014). Given a textual relation  $t_i$  and its embedding  $h_m$ , the predicted conditional probability of a KB relation  $r_j$  is thus:

$$p(r_j|t_i) = \text{softmax}(W_o h_o + b_o)_j, \quad (3)$$

where  $(\cdot)_j$  denotes the  $j$ -th element of a vector, and  $h_o$  is the state of the output GRU cell:

$$h_o = \text{GRU}(\phi(<GO>), h_m), \quad (4)$$

where  $<GO>$  is a special token indicating the start of decoding. The training objective is to minimize

$$\Theta = \frac{1}{|\mathcal{E}|} \sum_{i,j:\tilde{p}(r_j|t_i)>0} (\log p(r_j|t_i) - \log \tilde{p}(r_j|t_i))^2, \quad (5)$$

where  $\mathcal{E}$  is the edge set of the relation graph. It is modeled as a regression problem, similar to GloVe (Pennington et al., 2014).

**Baseline.** We also define a baseline approach where the unnormalized co-occurrence counts are directly used. The objective is to maximize:

$$\Theta' = \frac{1}{\sum_{i,j} n_{ij}} \sum_{i,j:n_{ij}>0} n_{ij} \log p(r_j|t_i). \quad (6)$$

It also corresponds to local statistics based embedding, i.e., when the embedding model is trained on individual occurrences of relational facts with distant supervision. Therefore, we call it **Local Relation Embedding (LoRE)**.

## 5 Augmenting Relation Extraction

Learned from global co-occurrence statistics of relations, our approach provides semantic matching information of textual and KB relations, which is often complementary to the information captured by existing relation extraction models. In this section we discuss how to combine them together to achieve better relation extraction performance.

We follow the setting of distantly supervised relation extraction. Given a text corpus and a KB with relation set  $\mathcal{R}$ , the goal is to find new relational facts from the text corpus that are not already contained in the KB. More formally, for each entity pair  $(e, e')$  and a set of *contextual sentences*  $C$  containing this entity pair, a relation extraction model assigns a score  $E(z|C)$  to each candidate relational fact  $z = (e, r, e'), r \in \mathcal{R}$ . On the other hand, our textual relation embedding model works on the sentence level. It assigns a score  $G(z|s)$  to each contextual sentence  $s$  in  $C$  as for how well the textual relation  $t$  between the entity pair in the sentence matches the KB relation  $r$ , i.e.,  $G(z|s) = p(r|t)$ . It poses a challenge to aggregate the sentence-level scores to get a set-level score  $G(z|C)$ , which can be used to combine with the original score  $E(z|C)$  to get a better evaluation of the candidate relational fact.

One straightforward aggregation is max pooling, i.e., only using the largest score  $\max_{s \in C} G(z|s)$ , similar to the at-least-one strategy used by Zeng et al. (2015). But it will lose the useful signals from those neglected sentences (Lin et al., 2016). Because of the wrong labeling problem, mean pooling is problematic as well. The wrongly labeled contextual sentences

tend to make the aggregate scores more evenly distributed and therefore become less informative. The number of contextual sentences positively supporting a relational fact is also an important signal, but is lost in mean pooling.

Instead, we use summation with a trainable  $cap$ :

$$G(z|C) = \min(cap, \sum_{s \in C} G(z|s)), \quad (7)$$

In other words, we additively aggregate the signals from all the contextual sentences, but only to a bounded degree.

We simply use a weighted sum to combine  $E(z|C)$  and  $G(z|C)$ , where the trainable weights will also handle the possibly different scale of scores generated by different models:

$$\tilde{E}(z|C) = w_1 E(z|C) + w_2 G(z|C). \quad (8)$$

The original score  $E(z|C)$  is then replaced by the new score  $\tilde{E}(z|C)$ . To find the optimal values for  $w_1$ ,  $w_2$  and  $cap$ , we define a hinge loss:

$$\Theta_{Merge} = \frac{1}{K} \sum_{k=1}^K \max\{0, 1 + \tilde{E}(z_k^-) - \tilde{E}(z_k^+)\}, \quad (9)$$

where  $\{z_k^+\}_{k=1}^K$  are the true relational facts from the KB, and  $\{z_k^-\}_{k=1}^K$  are false relational facts generated by replacing the KB relation in true relational facts with incorrect KB relations.

## 6 Experiments

In this experimental study, we demonstrate the effectiveness of GloRE by showing that it could significantly improve the performance of several recent relation extraction methods.

### 6.1 Experimental Setup

**Dataset.** Following the literature (Hoffmann et al., 2011; Surdeanu et al., 2012; Zeng et al., 2015; Lin et al., 2016), we use the relation extraction dataset introduced in (Riedel et al., 2010), which was generated by aligning New York Times (NYT) articles with Freebase (Bollacker et al., 2008). Articles from year 2005-2006 are used as training, and articles from 2007 are used as testing. Some statistics are listed in Table 1. There are 53 target KB relations, including a special relation NA indicating that there is no target relation between an entity pair.

Data	# of sentences	# of entity pairs	# of relational facts from KB
Train	570,088	291,699	19,429
Test	172,448	96,678	1,950

Table 1: Statistics of the NYT dataset.

We follow the approach described in Section 3 to construct the relation graph from the NYT training data. The constructed relation graph contains 321,447 edges with non-zero weight. We further obtain a training set and a validation set from the edges of the relation graph. We have observed that using a validation set totally disjoint from the training set leads to unstable validation loss, so we randomly sample 300K edges as the training set, and another 60K as the validation set. The two sets can have some overlap. For the merging model (Eq. (9)), 10% of the edges are reserved as the validation set.

**Relation extraction models.** We evaluate with four recent relation extraction models whose source code is publicly available<sup>3</sup>. We use the optimized parameters provided by the authors.

- **CNN+ONE** and **PCNN+ONE** (Zeng et al., 2015): A convolutional neural network (CNN) is used to embed contextual sentences for relation classification. Multi-instance learning with at-least-one (ONE) assumption is used to combat the wrong labeling problem. In PCNN, piecewise max pooling is used to handle the three pieces of a contextual sentence (split by the two entities) separately.
- **CNN+ATT** and **PCNN+ATT** (Lin et al., 2016): Different from the at-least-one assumption which loses information in the neglected sentences, these models learn soft attention weights (ATT) over contextual sentences and thus can use the information of all the contextual sentences. PCNN+ATT is the best-performing model on the NYT dataset.

**Evaluation settings and metrics.** Similar to previous work (Riedel et al., 2010; Zeng et al., 2015), we use two settings for evaluation: (1) Held-out evaluation, where a subset of relational facts in KB is held out from training (Table 1), and is later used to compare against newly discovered relational facts. This setting avoids human labor but can introduce some false negatives because of the

<sup>3</sup><https://github.com/thunlp/NRE>

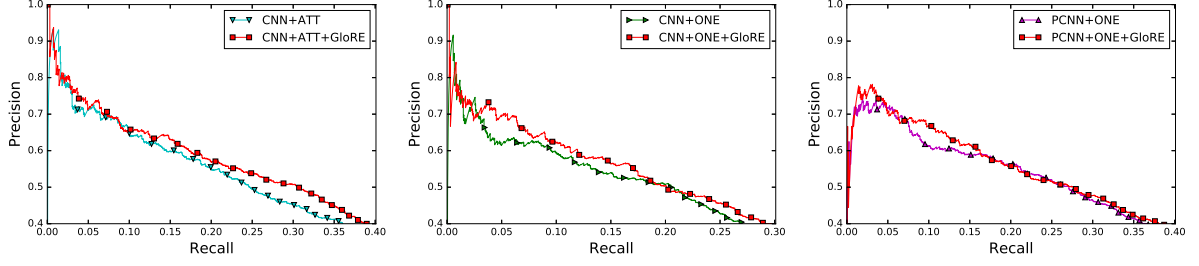


Figure 4: Held-out evaluation: other base relation extraction models and the improved versions when augmented with GloRE.

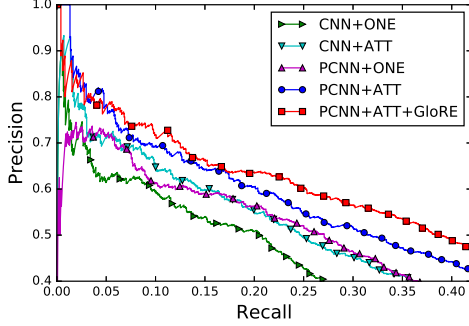


Figure 5: Held-out evaluation: the previous best-performing model can be further improved when augmented with GloRE.

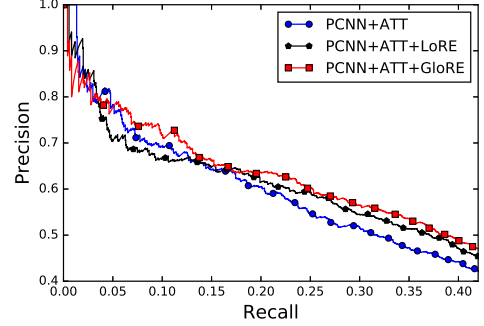


Figure 6: Held-out evaluation: LoRE vs. GloRE.

incompleteness of the KB. (2) Manual evaluation, where the discovered relational facts are manually judged by human experts. For held-out evaluation, we report the precision-recall curve. For manual evaluation, we report  $Precision@N$ , i.e., the precision of the top  $N$  discovered relational facts.

**Parameter settings.** Hyper-parameters of our model are selected based on the validation set. For the embedding model, the mini-batch size is set to 128, and the state size of the GRU cells is 300. For the merging model, the mini-batch size is set to 1024. We use Adam (Kingma and Ba, 2014) with parameters suggested by the authors for optimization. Word embeddings are initialized with the 300-dimensional word2vec (Mikolov et al., 2013) vectors pre-trained on the Google News corpus<sup>4</sup>. Early stopping based on the validation set is employed. Our model is implemented using Tensorflow (Abadi et al., 2016), and the source code is available at <https://github.com/ppuliu/GloRE>.

## 6.2 Held-out Evaluation

**Existing Models + GloRE.** We first show that our approach, GloRE, can improve the performance of the previous best-performing model,

<sup>4</sup><https://code.google.com/archive/p/word2vec/>

Precision@ $N$	100	300	500	700	900	1000
PCNN+ATT	<b>97.0</b>	93.7	92.8	89.1	85.2	83.9
PCNN+ATT+LoRE	<b>97.0</b>	95.0	94.2	91.6	89.6	87.0
PCNN+ATT+GloRE	<b>97.0</b>	<b>97.3</b>	<b>94.6</b>	<b>93.3</b>	<b>90.1</b>	<b>89.3</b>

Table 2: Manual evaluation: false negatives from held-out evaluation are manually corrected by human experts.

PCNN+ATT, leading to a new state of the art on the NYT dataset. As shown in Figure 5, when PCNN+ATT is augmented with GloRE, a consistent improvement along the precision-recall curve is observed. It’s worth noting that although PCNN+ATT+GloRE seems to be inferior to PCNN+ATT when recall  $< 0.05$ , as we will show via manual evaluation, it is actually due to false negatives.

We also show in Figure 4 that the improvement brought by GloRE is general, not just specific to PCNN+ATT; the other three models also get a consistent improvement when augmented with GloRE. The results suggest that our textual relation embedding approach with global statistics indeed captures useful information for relation extraction that is not captured by these sentence embedding based models.

**LoRE v.s. GloRE.** We compare GloRE with the baseline approach LoRE (Section 4) to show the advantage of normalization on global statistics. We use PCNN+ATT as the base relation extraction model. As shown in Figure 6, GloRE consis-

Contextual Sentence	Textual Relation	PCNN+ATT Predictions	LoRE Predictions	GloRE Predictions
[Alfred Blumstein] <sub>head</sub> , a criminologist at [Carnegie Mellon University] <sub>tail</sub> , called ...	$\xleftarrow{\text{appos}}$ criminologist $\xrightarrow{\text{nmod:at}}$	NA (0.63) <b>employee_of</b> (0.36) founder_of (0.00)	<b>employee_of</b> (1.00) NA (0.00) founder_of (0.00)	<b>employee_of</b> (0.96) NA (0.02) founder_of (0.02)
[Langston Hughes] <sub>head</sub> , the American poet, playwright and novelist, came to [Spain] <sub>tail</sub> to ...	$\xleftarrow{\text{-nsubj}}$ came $\xrightarrow{\text{to}}$	NA (0.58) nationality (0.38) place_lived (0.01)	place_of_death (0.35) <b>NA</b> (0.33) nationality (0.21)	<b>NA</b> (0.73) contain_location (0.07) employee_of (0.06)

Table 3: Case studies. We select entity pairs that have only one contextual sentence, and the head and tail entities are marked. The top 3 predictions from each model with the associated probabilities are listed, with the correct relation bold-faced.

tently outperforms LoRE. It is worth noting that LoRE can still improve the base relation extraction model when recall  $> 0.15$ , further confirming the usefulness of directly embedding textual relations in addition to sentences.

### 6.3 Manual Evaluation

Due to the incompleteness of the knowledge base, held-out evaluation introduces some false negatives. The precision from held-out evaluation is therefore a lower bound of the true precision. To get a more accurate evaluation of model performance, we have human experts to manually check the false relational facts judged by held-out evaluation in the top 1,000 predictions of three models: PCNN+ATT, PCNN+ATT+LoRE, and PCNN+ATT+GloRE, and report the corrected results in Table 2. Under manual evaluation, PCNN+ATT+GloRE achieves the best performance in the full range of  $N$ . In particular, for the top 1,000 predictions, GloRE improves the precision of the previous best model PCNN+ATT from 83.9% to 89.3%. The manual evaluation results reinforce the previous observations from held-out evaluation.

### 6.4 Case Study

Table 3 shows two examples. For better illustration, we choose entity pairs that have only one contextual sentence.

For the first example, PCNN+ATT predicts that most likely there is no KB relation between the entity pair, while both LoRE and GloRE identify the correct relation with high confidence. The textual relation clearly indicates that the head entity is (appos) a criminologist at (nmod:at) the tail entity.

For the second example, there is no KB relation between the entity pair, and PCNN+ATT is indeed able to rank NA at the top. However, it is still quite confused by *nationality*, probably because it has learned that sentences about a person and a country with many words about

profession (“poet,” “playwright,” and “novelist”) likely express the person’s nationality. As a result, its prediction on NA is not very confident. On the other hand, GloRE learns that if a person “came to” a place, likely it is not his/her birthplace. In the training data, due to the wrong labeling problem of distant supervision, the textual relation is wrongly labeled with *place\_of\_death* and *nationality* a couple of times, and both PCNN+ATT and LoRE suffer from the training noise. Taking advantage of global statistics, GloRE is more robust to such noise introduced by the wrong labeling problem.

## 7 Conclusion

Our results show that textual relation embedding based on global co-occurrence statistics with KB relations captures useful information for relation extraction, and, as a result, can improve existing relation extraction models. Large-scale training data for embedding can be easily solicited from distant supervision, and the global statistics of relations provide a natural way to combat the wrong labeling problem of distant supervision.

The idea of relation embedding based on global statistics can be further expanded along several directions. In this work we have focused on embedding textual relations, but it is in principle beneficial to jointly embed knowledge base (KB) relations as well as entities. Recently a joint embedding approach has been attempted in the context of knowledge base completion (Toutanova et al., 2015), but it is still based on local statistics, i.e., individual relational facts. Joint embedding with global statistics remains an open problem. On the other hand, we have analyzed the distribution difference of words and textual relations, and its impact on normalizing the co-occurrence statistics of relations. A more thorough comparative study of normalization strategies can shed light on future use of co-occurrence statistics of relations.



## References

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*.
- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International conference on Management of data*. ACM, pages 1247–1250.
- John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods* 39(3):510–526.
- Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 724–731.
- Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 740–750.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv:1406.1078*.
- Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, page 423.
- Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6):391.
- Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89(1):31–71.
- Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 427–434.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 541–550.
- Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL on Interactive poster and demonstration sessions*. Association for Computational Linguistics, page 22.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.
- Rémi Lebrete and Ronan Collobert. 2014. Word embeddings through hellinger PCA. *European Chapter of the Association for Computational Linguistics* page 482.
- Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 2124–2133.
- Yang Liu, Sujian Li, Furu Wei, and Heng Ji. 2016. Relation classification via modeling augmented dependency paths. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 24(9):1585–1594.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. pages 3111–3119.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1003–1011.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1532–1543.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pages 148–163.
- Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

- Douglas LT Rohde, Laura M Gonnerman, and David C Plaut. 2005. An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM* 8:627–633.
- Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 1201–1211.
- Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 455–465.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. pages 3104–3112.
- Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoi-fung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1499–1509.
- Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.
- Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015a. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv:1506.07650*.
- Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv:1601.03651*.
- Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015b. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1785–1794.
- Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research* 3(Feb):1083–1106.
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1753–1762.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the International Conference on Computational Linguistics*. pages 2335–2344.
- Min Zhang, Jie Zhang, and Jian Su. 2006. Exploring syntactic features for relation extraction using a convolution tree kernel. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, pages 288–295.

## A Word and Relation Distribution

Because textual relations are sequences composed of words and dependency relations, they are less likely to collide than words. As a result, most of textual relations only occur a few times. From the New York Times corpus, we collect the occurrence and co-occurrence statistics of words and textual relations. Two words co-occur if they appear in the same sentence. The co-occurring relationship of textual and KB relations are established via distant supervision, as described in Section 3. As can be seen, the relation distributions are more skewed towards the lower end of the x-axis than the word distributions. Very few textual relations occur more than 100 times. Rare relations are the norm, not the exception. But it is worth mentioning that the most frequent textual relation still occurs more than 40 thousand times (not shown in the figures). Normalization of relation co-occurrence counts shall consider the characteristics of relation distributions, and give proper weights to rare relations.

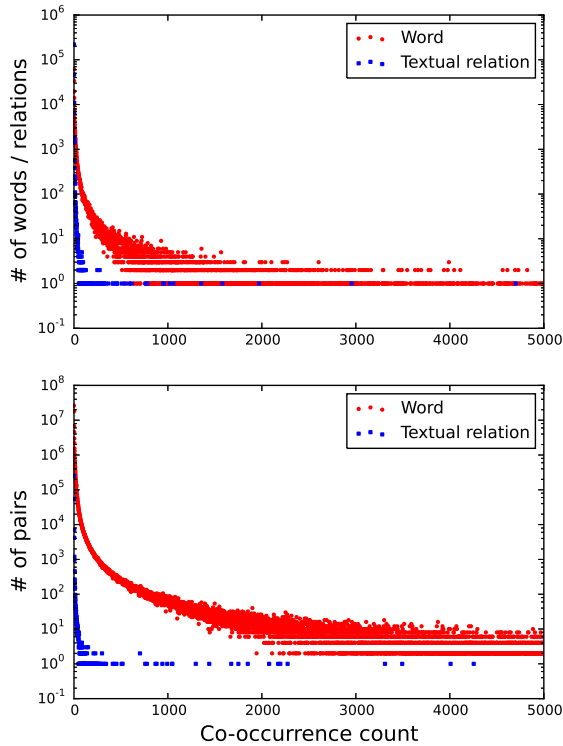


Figure 7: *Top*: Word and relation frequency distribution. The x-axis is the number of occurrences of a word or a textual relation, and the y-axis is the number of words or textual relations with a certain number of occurrences. *Bottom*: Word and relation co-occurrence distribution. The x-axis is the co-occurrence count of a word pair or a textual-KB relation pair, and the y-axis is the number of pairs with a certain number of co-occurrences.